

DOCUMENT RESUME

ED 416 213

TM 028 065

AUTHOR Woldbeck, Tanya
TITLE A Primer on Logistic Regression.
PUB DATE 1998-01-23
NOTE 22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January 23, 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavioral Science Research; *Chi Square; Computer Software; Goodness of Fit; Item Bias; *Regression (Statistics); Test Items
IDENTIFIERS *Logistic Regression; *Odds Ratio

ABSTRACT

This paper introduces logistic regression as a viable alternative when the researcher is faced with variables that are not continuous. If one is to use simple regression, the dependent variable must be measured on a continuous scale. In the behavioral sciences, it may not always be appropriate or possible to have a measured dependent variable on a continuous scale. Logistic regression, a technique derived from logit modeling or logit analysis, has been the analysis of choice in many areas of research in this situation. Logistic regression, like the other regression analyses still looks at the relationship between the variables of interest as the core focus of the analysis, but it uses the concept of the odds ratio as its measure of association. An example illustrates the interpretation of coefficients using the odds ratio. Traditionally the chi-square statistic has been used to test for independence between variables. In logistic regression, the likelihood-ratio-chi-squared statistic (G-squared) is important in comparing the observed and expected frequencies of the variable of interest to assess the goodness of fit of the model. Logistic regression has been used in many areas of research in the behavioral sciences, and can be used in detecting differential item functioning. It is a part of statistical software packages, and an example is included of use of this viable and efficient tool for statistical analysis. (Contains 5 tables, 1 figure, and 18 references.) (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document.. *

Running Head: LOGISTIC REGRESSION

A Primer on Logistic Regression

Tanya Woldbeck

Texas A&M University 77843-4225

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

T. WOLDBECK

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to
improve reproduction quality.

- Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

Paper presented at the annual meeting of the Southwest
Educational Research Association, Houston, TX, January 23, 1998.

Abstract

With increasing emphasis being placed on one's choice of statistical analyses, it is important to understand the decisions you make in research design and choice of analysis, and then how to use them and what they mean. The following paper introduces the reader to logistic regression as a viable alternative when faced with dependent variables that are not continuous.

One of the most controversial topics in the area of statistics and design today seems to be the argument over statistical significance. An entire volume of the Journal of Experimental Education (1993) was devoted to the topic. Within the issue, Carver (1993) reminds us to consider the effects of our sample size on the statistical significance of our results, and Snyder and Lawson (1993) encourage us to use effect sizes to support our decisions about result importance. In addition, it would seem important, when designing a study, that the researcher select the appropriate statistical analysis as well. That is, effect sizes computed from an incorrect analysis may be just as inappropriate as not computing effect sizes from a correct analysis.

The current paper elaborates the statistical method known as logistic regression. The appropriate use of logistic regression will be discussed and examples of its application will be explored. In addition, the use of logistic regression in certain areas will be compared and contrasted with other analyses, specifically with linear regression and discriminant analysis. While an in-depth analysis of logit modeling and logistic regression is beyond the scope of the present effort, a useful introduction will be offered.

A look at logistic regression as an option in research should begin with a look at conventional, non-logistic regression and its assumptions. It is the breakdown of these assumptions where logistic regression comes to bear on the choice of appropriate statistical analyses.

According to Pedhazur and Schmelkin's (1991) coverage of

regression, the first assumption is that the proper regression equation, and therefore, coefficients, are used to correctly summarize the independent variable's effect on the dependent variable. A second assumption states the importance of correct measurement of the variables in question. A final assumption deals specifically with error terms and includes the assumption of homoscedasticity, the assumption that error terms, or residuals, are uncorrelated and that residuals are normally distributed (p. 389).

In a more extensive coverage of simple regression, Schroeder, Sjoquist, and Stephan (1986) also remind us that if one is to use simple regression, the dependent variable must be measured on a continuous scale. Haase and Thompson (1992) caution against dichotomizing naturally continuous variables, which is why the field has moved towards the use of regression techniques in the first place (Edington, 1964, 1974; Elmore & Woehlke, 1988). However, in the behavioral sciences, it may not always be appropriate, or possible, to have a measured dependent variable on a continuous scale. Responses on a dependent variable may inherently be categorical or dichotomous in nature. Such a categorical dependent variable leads certain assumptions of simple regression to fall apart, and regression analysis, therefore, becomes meaningless. As Schroeder et al. (1986) noted, "While [the use of 0-1 dummy variables] are appropriate as explanatory variables, ordinary least squares (OLS) regression analysis is not appropriate when a 0-1 or other limited choice variable is the

dependent variable" (p. 79).

The problem of using OLS regression with a dichotomous dependent variable begins with a look at the regression equation itself. Consider the example of being blonde or not being blonde. The dependent variable (B) would be equal to one if you were a blonde and zero if you were not a blonde. For simplicity, let us assume the independent variable (F) is a score on a measure called "Do you have fun"? Now consider the regression equation, $B = \alpha + \beta F$. Using simple regression techniques, it could be possible to obtain values for B that are greater than one, less than zero, and in between. These values, given that B is a dichotomous, not a continuous variable, simply do not make sense as estimates. It would be fair to say that assumption one of simple regression would be violated, and results would be meaningless. This will be discussed later in greater detail.

Further problems evident with the use of categorical dependent variables involve the assumptions surrounding the residuals. Schroeder et al. (1986) state that, "[T]he variability of residuals obtained from [the simple regression equation] will depend on the size of the independent variable, suggesting that heteroscedasticity is a problem..." (pp. 79-80). In addition, residual terms with a dichotomous variable are certainly not normally distributed.

So now the question is, given a dichotomous or categorical dependent variable, how can we effectively summarize the relationship between an independent and a dependent variable?

Techniques which can be used include discriminant analysis (DeMaris, 1992), probit analysis, tobit analysis, and econometrics (Schroeder et al., 1986). In the past 20 years or so, logistic regression, a technique derived from logit modeling or logit analysis, has been the analysis of choice in various areas of research (French & Miller, 1996; Morgan & Teachman, 1988; Press & Wilson, 1978; Swaminathan & Rogers, 1990). This technique has especially useful in the area of epidemiology (Lemeshow & Hosmer, 1982) where the dependent variable is often categorical (e.g., heart attack or no heart attack).

To summarize thus far, we know that logistic regression is a viable statistical alternative when addressing a dependent variable which is dichotomous or categorical in nature. DeMaris (1992) posits that the "advent of loglinear modeling has revolutionized the multivariate analysis of categorical data" (p. 1)...and nicely summarizes the effect of a given predictor on a dependent variable in a "compact and elegant" manner (p. 2). Before an explanation of logistic regression can begin, however, it is "important to understand that the goal of analysis using this method is the same as that of any model-building technique used in statistics: To find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome... and a set of independent... variables" (Hosmer & Lemeshow, 1989, p. 1).

To better understand the fundamentals of logistic regression, similarities and contrasts with linear regression will again be used along with an extension of the blonde example mentioned above.

In addition, for heuristic purposes, logistic regression will be contrasted with both linear regression and discriminant through an example later in the paper.

In this simplified hair color example, the independent variable will continue to be one's score on a survey called "Do You Have Fun?". Scores on the survey are continuous in nature and range from 0 (little to no fun) to 20(all the time). Scores and outcomes are represented in Table 1.

Insert Table 1 about here

In the table, a score of 0 on the dependent variable "hair" represents a person with non-blond hair color, and a score of 1 represents a person with blond hair. This zero, one coding is called dummy coding (Rice, 1994). Design coding can also be used which assigns a one and a negative one to the categories. Rice (1994) outlines these various coding schemes.

The question still remains, remember, "is there a relationship between hair color, and how much fun a person has?" A scatterplot of this data may look something like what you see in Figure 1. The dichotomous nature of the data is evident, but a relationship is not very clear.

Insert Figure 1 about here

"The first difference [between linear and logistic regression]

concerns the nature of the relationship between the outcome and the independent variable" (Hosmer & Lemeshow, 1989, p. 5). We are still concerned with the expected value of F (DV) given a value of B (IV) (called the conditional mean). Now, however, with dichotomous data, the conditional mean must fall between zero and positive one. The distribution used is the logistic distribution. In logistic regression, $\Pi(x)$ represents the conditional mean of the dependent variable given the independent variable. For mathematical simplicity, $\Pi(x)$ can be transformed to $g(x)$ through mathematical manipulation. $g(x)$ has many desirable properties to warrant this logit transformation. $g(x)$ is linear, and may be continuous (Hosmer & Lemeshow, p. 7). The transformation forces the probabilities (dependent variable outcome) to fall between 0 and 1, a desired outcome with this type of variable (Rice, 1994). This same logit transformation is used in Item Response Theory (IRT) as a method of placing ability and item characteristics on the same scale. In fact, the Rasch model of IRT is a form of logistic regression (Hambleton & Swaminathan, 1985).

A second difference has to do with the error term. You will recall, from the simple regression assumptions listed above, that residuals in linear regression are normally distributed. This is no longer the case with dichotomous data. Error (ε) takes on one of two possible variables, $\varepsilon = -\pi(x)$ when the dependent variable is zero, and probability $1-\pi(x)$ and $\varepsilon = 1-\pi(x)$ when the dependent variable is 1 and probability $\pi(x)$.

As in regression, it would first be appropriate to estimate

our unknown parameters. In linear regression, it is the method called maximum likelihood that "yields values for the unknown parameters which maximize the probability of obtaining the observed set of data" (Hosmer & Lemeshow, 1989, p. 8). The probability of the obtained data is termed the likelihood function. In logistic regression, for mathematical simplicity, logarithmic manipulations are invoked to produce the log likelihood. This new equation becomes the basis for estimating our unknown parameters, our regression coefficients (Hosmer & Lemeshow, 1989).

Once we have our coefficients, it is time to ask: "What do the estimated coefficients in the model tell us about the research questions that motivated the study?" (Hosmer & Lemeshow, 1989, p. 38). In linear regression, the slope (β) tells us the change in the dependent variable given a one unit change in the independent variable. In logistic regression, the logit transformation $g(x)$ must be used and the slope coefficient "represents the change in the logit for a change in one unit in the independent variable" (Hosmer & Lemeshow, 1989, p. 39). Further interpretation of the coefficients depends on the nature of the independent variable.

Logistic regression, like other regression analyses, still looks at the relationship between variables of interest as the core focus of analysis. However, logistic regression uses the concept of the odds ratio as its measure of association. Odds are the ratio of events to nonevents, and odds ratios can be considered a ratio of two odds (Morgan & Teachman, 1988). DeMaris summarizes by stating that, "In categorical data analysis, the 'effect' of one

variable upon another is best expressed in terms of odds ratios" (p. 6).

An example may further understanding. Let us use the blonde example. To make both variables dichotomous, we will categorize our continuous variable labeled "FUN" in table one. Many (Haase & Thompson, 1992; Hosmer & Lemeshow, 1989) have warned against categorizing variables because it discards variation which is the basis of all analyses. However, for heuristic purposes, we shall code any score below 10 on the "Do You Have Fun?" survey as 0 and any score of 10 and above will receive a 1 coding. The independent ratio will be represented by the letter F and the dependent variable will be represented by the letter B. Using the odds ratio as the basis for a measure of association, the interpretation of coefficients is as follows.

The odds of the outcome (B) being present ($B=1$) among individuals with $F=1$ is defined as $\pi(1)/[1-\pi(1)]$. Using the same logic, the odds of the outcome being present ($B=1$) among individuals $F=0$ is $\pi(0)/[1-\pi(0)]$. The next step is to find the odds ratio (ψ) which is the ratio of the odds for $F=1$ and $F=0$. The log of the odds ratio (the log-odds) is then computed, and with a dichotomous independent variable, the value of the log odds is found to be equal to β , our regression coefficient. The odds ratio then (ψ) approximates how much more likely (or unlikely) it is for the outcome to be present among those with $F=1$ than among those with $F=0$ (Hosmer & Lemeshow, 1989). In our example, if $\psi=2$, we could say that blondes are twice as likely to have fun than non-

blondes.

Once we have our coefficients, and we are able to interpret them, we are ready to assess "the fit of an estimated logistic regression model with the assumption that we are at least preliminarily satisfied with our efforts at the model building stage" (Hosmer & Lemeshow, 1989, p. 135). Traditionally, the Pearson Chi-squared (χ -squared) statistic has been used to test for independence between variables. Rejecting the null hypothesis in this case would lend support to having confidence in the fact that the two variables were associated in some way (DeMaris, 1992).

Another chi-squared statistic called the likelihood-ratio chi-squared statistic (G-squared) is important in logistic regression. G-squared is used to compare observed and expected frequencies of the variables of interest to assess the goodness of fit for the model. The approach used is testing for independence between observed and expected data. A small test statistic would lead us to not reject the null hypothesis and we can then have confidence that our model fits the data well (DeMaris, 1992). Lemeshow and Hosmer (1982) outline a number of fit statistics including a technique they came up with.

A technique that has often been used in statistics and design is discriminant analysis (see Klecka, 1980). This would also be a viable alternative to use with dichotomous dependent variables. Predictive discriminant analysis (PDA) is used to predict group membership (dichotomous dependent variable) with certain independent continuous variables (Klecka, 1980). Press and Wilson

(1978) posit, however, that logistic regression and the use of maximum likelihood estimators (MLEs) are generally superior to discriminant function estimators for several reasons. Press and Wilson claim that logistic regression is more robust; "i.e., many types of underlying assumptions lead to the same logistic formulation" (p. 700). These authors also feel that MLEs are more consistent, more efficient, and that "[t]he logistic regression model is well-known to have sufficient statistics associated with it" (Press & Wilson, 1978, p. 701). In addition, Rice (1994) points out that the assumption of multivariate normality necessary for PDA does not hold up with dichotomous, categorical dependent variables. As mentioned earlier, the predicted values (probabilities in the case of PDA) need to range within zero and positive one, an outcome only possible with logistic regression.

Logistic regression has found great utility in many areas of research in the behavioral sciences. Swaminathan and Rogers (1990) have used logistic regression in detecting differential item functioning (DIF). DIF is the study of test items and how they function with various test-takers. Logistic regression has been used in research ranging from marriage and family therapy (Morgan & Teachman, 1988) to graduate student persistence with financial aid as the predicting variable (Murdock & Nix-Mayer, 1995).

Logistic regression is a fast part of all statistical software packages, and an example is included at the end of the paper. The example shows that similar results are achieved when one runs a discriminant analysis and logistic regression; comparable aspects

of the results of each have been bolded to facilitate this comparison. A conventional, non-logistic analysis is also included. The logistic regression, conventional regression, and discriminant analyses can be found in Tables 2, 3, and 4, respectively.

 Insert Tables 2, 3, and 4 about here

For reasons stated above, then, logistic regression seems the better choice of the two. The example also shows that linear regression is not appropriate for these types of data, given the dichotomous nature of the dependent variable. The data for the SPSS run (a statistical software package which includes all of the aforementioned analyses) were taken from Holzinger and Swineford (1939).

In sum, logistic regression, while not too extensively utilized, seems to be a viable and very efficient statistical tool in the area of statistical analysis when the researcher is left with dichotomous variables. The technique is not mentioned in any of the tabulation articles looking at the use of statistics in major journals (Edington, 1964, 1974; Elmore & Woehlke, 1988). While the mathematical components sometimes are overwhelming, statistical software packages help in that area by simplifying the math. What is left then is a powerful tool for use with dichotomous dependent variables, which are often encountered in the behavioral sciences.

References

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

DeMaris, A. (1986). Logit modeling: practical applications. Newbury Park, CA: Sage.

Edington, E.S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.

Edington, E.S. (1974). A new tabulation of statistical procedures in APA journals. American Psychologist, 29(1), 25-26.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

French, A.W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. Journal of Educational Measurement, 33(3), 315-332.

Hosmer, D.W., & Lemeshow, S. (1989). Applied logistic regression. New York: John Wiley & Sons.

Klecka, W.R. (1980). Discriminant analysis. Newbury Park, CA: Sage.

Lemeshow, S. & Hosmer, D.W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology, 115(1), 92-106.

Morgan, S.P., & Teachman, J.D. (1988). Logistic regression:

Description, examples, and comparisons. Journal of Marriage and the Family, 50, 929-936.

Murdock, T., & Nix-Mayer, L. (1995, May). The effect of types of financial aid on student persistence towards graduation. Paper presented at the 35th annual forum of the Association for Instructional Research, Boston, MA.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Earlbaum.

Press, J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, 73, 699-705.

Rice, J.C. (1994). Logistic regression: An introduction. In B.Thompson (Ed.), Advances in social science methodology (Vol.3, pp. 3-28). Greenwich, CT: JAI Press.

Schroeder, L.D., Sjoquist, D.L., & Stephan, P.E. (1986). Understanding regression analysis: An introductory guide. Newbury Park, CA: Sage.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61(4), 334-349.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Thompson, B. (Ed.). (1993). Statistical significance testing in contemporary practice: Some proposed alternatives with

comments from journal editors. Journal of Experimental Education,
61(4).

Table 1

Score on "Do You Have Fun?" Survey and Hair Color

ID	FUN	HAIR
1	5	0
2	16	1
3	15	1
4	7	0
5	6	0
6	14	1
7	19	1
8	4	0
9	8	0
10	17	1

Table 2

Sample Output Using Logistic Regression

LOGISTIC REGRESSION VAR=grade
 /METHOD=ENTER t10 t15 t22 t6
 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

Dependent Variable.. GRADE
 Beginning Block Number 0. Initial Log Likelihood Function
 -2 Log Likelihood 416.71297
 * Constant is included in the model.
 Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number
 1.. T10 SPEEDED ADDITION TEST
 T15 MEMORY OF TARGET NUMBERS
 T22 MATH WORD PROBLEM REASONING
 T6 PARAGRAPH COMPREHENSION TEST

Estimation terminated at iteration number 3 because
 Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood 368.330
 Goodness of Fit 297.455
 Cox & Snell - R² .148
 Nagelkerke - R² .198

	Chi-Square	df	Significance
Model	48.383	4	.0000
Block	48.383	4	.0000
Step	48.383	4	.0000

----- Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
T10	.0301	.0056	28.8601	1	.0000	.2539	1.0305
T15	.0012	.0165	.0051	1	.9430	.0000	1.0012
T22	.0193	.0152	1.6141	1	.2039	.0000	1.0195
T6	.0821	.0408	4.0433	1	.0443	.0700	1.0855
Constant	-4.3545	1.5905	7.4960	1	.0062		

Table 3

Sample Output for Conventional, Non-logistic Regression

```

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT grade
/METHOD=ENTER t10 t15 t22 t6 .

```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
-------	---	----------	-------------------	----------------------------

1	<u>.386 (a)</u>	<u>.149</u>	.137	.46
---	-----------------	-------------	------	-----

a Predictors: (Constant), T6, T15, T10, T22

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.170	4	2.793	12.928	.000 (a)
	Residual	63.940	296	.216		
	Total	75.110	300			

a Predictors: (Constant), T6, T15, T10, T22

b Dependent Variable: GRADE

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		sig.
		B	SE	Beta	t	
1	(Constant)	6.591	.325		20.252	.000
	T10	6.440E-03	.001	.323	5.893	.000
	T15	3.743E-05	.004	.001	.011	.991
	T22	4.041E-03	.003	.074	1.237	.217
	T6	1.718E-02	.009	.120	1.973	.049

a Dependent Variable: GRADE

Table 4

Sample Output For Discriminant Analysis

DISCRIMINANT
 /GROUPS=grade(7 8)
 /VARIABLES=t10 t15 t22 t6
 /ANALYSIS ALL
 /PRIORS EQUAL
 /CLASSIFY=NONMISSING POOLED .

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.175(a)	100.0	100.0	<u>.386</u>

a First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.851	47.820	4	.000

Standardized Canonical Discriminant Function Coefficients

	Function 1
T10	.850
T15	.002
T22	.206
T6	.329

Structure Matrix

	Function 1
T10	.890
T6	.512
T22	.364
T15	.119

Figure 1
Scatterplot of FUN score by HAIR Color

